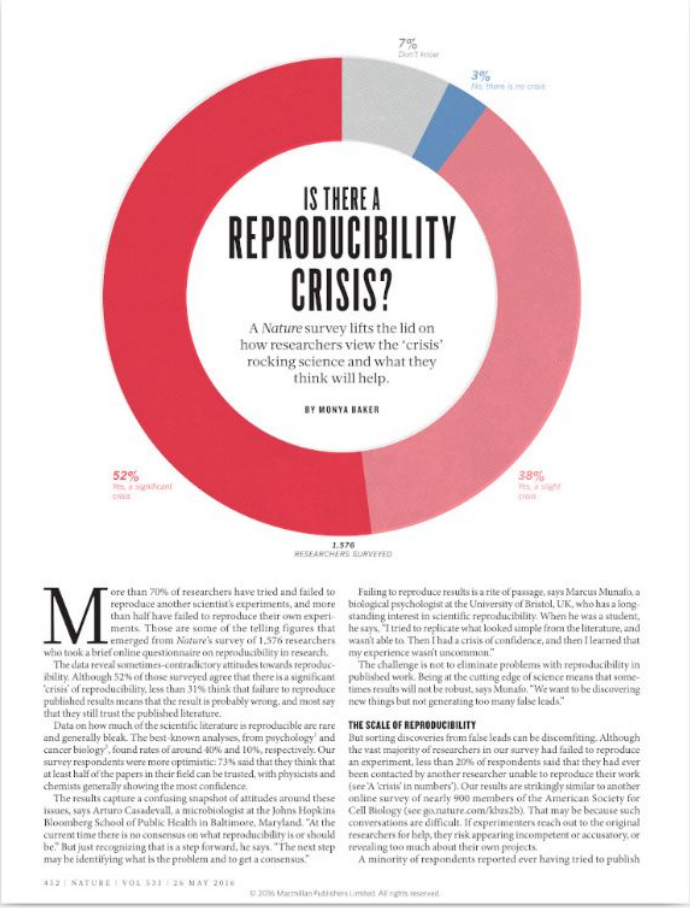
Environments for reproducibility

Javier Moldón
IAA-CSIC (Granada)

Herramientas para la reproducibilidad del análisis científico
Curso del CSIC
April 20, 2022

“Reproducibility crisis” in the sciences? (Baker 2016, Nature 533, 452)



IS THERE A REPRODUCIBILITY CRISIS?

A *Nature* survey lifts the lid on how researchers view the 'crisis' rocking science and what they think will help.

BY MONYA BAKER

More than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments. Those are some of the telling figures that emerged from *Nature*'s survey of 1,576 researchers who took a brief online questionnaire on reproducibility in research.

The data reveal sometimes-contradictory attitudes towards reproducibility. Although 52% of those surveyed agree that there is a significant 'crisis' of reproducibility, less than 31% think that failure to reproduce published results means that the result is probably wrong, and most say that they still trust the published literature.

Data on how much of the scientific literature is reproducible are rare and generally bleak. The best-known analyses, from psychology[1] and cancer biology[2], found rates of around 40% and 10%, respectively. Our survey respondents were more optimistic: 73% said that they think that at least half of the papers in their field can be trusted, with physicists and chemists generally showing the most confidence.

The results capture a confusing snapshot of attitudes around these issues, says Arturo Casadevall, a microbiologist at the Johns Hopkins Bloomberg School of Public Health in Baltimore, Maryland. "At the current time there is no consensus on what reproducibility is or should be." But just recognizing that is a step forward, he says. "The next step may be identifying what is the problem and to get a consensus."

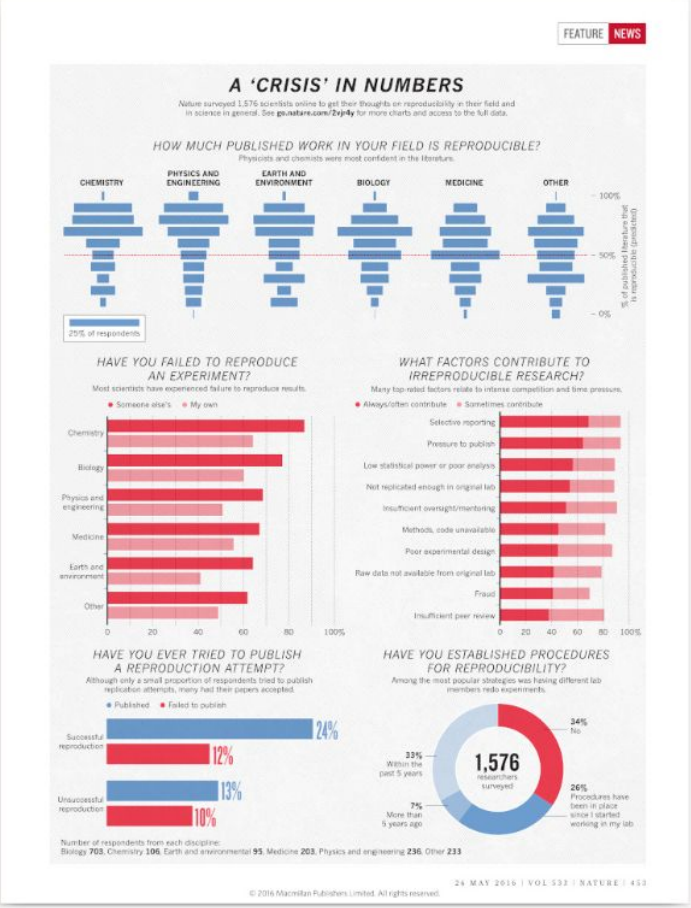
Failing to reproduce results is a rite of passage, says Marcus Munafo, a biological psychologist at the University of Bristol, UK, who has a long-standing interest in scientific reproducibility. When he was a student, he says, "I tried to replicate what looked simple from the literature, and wasn't able to. Then I had a crisis of confidence, and then I learned that my experience wasn't uncommon."

The challenge is not to eliminate problems with reproducibility in published work. Being at the cutting edge of science means that sometimes results will not be robust, says Munafo. "We want to be discovering new things but not generating too many false leads."

THE SCALE OF REPRODUCIBILITY

But sorting discoveries from false leads can be discomfiting. Although the vast majority of researchers in our survey had failed to reproduce an experiment, less than 20% of respondents said that they had ever been contacted by another researcher unable to reproduce their work (see 'A 'crisis' in numbers'). Our results are strikingly similar to another online survey of nearly 900 members of the American Society for Cell Biology (see go.nature.com/kbzs2b). That may be because such conversations are difficult. If experimenters reach out to the original researchers for help, they risk appearing incompetent or accusatory, or revealing too much about their own projects.

A minority of respondents reported ever having tried to publish



A 'CRISIS' IN NUMBERS

Nature surveyed 1,576 scientists online to get their thoughts on reproducibility in their field and in science in general. See **go.nature.com/2vjr4y** for more charts and access to the full data.

HOW MUCH PUBLISHED WORK IN YOUR FIELD IS REPRODUCIBLE?

Physicists and chemists were most confident in the literature.

HAVE YOU FAILED TO REPRODUCE AN EXPERIMENT?

Most scientists have experienced failure to reproduce results.

WHAT FACTORS CONTRIBUTE TO IRREPRODUCIBLE RESEARCH?

Many top-rated factors relate to intense competition and time pressure.

● Always/often contribute ● Sometimes contribute
Selective reporting
Pressure to publish
Low statistical power or poor analysis
Not replicated enough in original lab
Insufficient oversight/mentoring
Methods, code unavailable
Poor experimental design
Raw data not available from original lab
Fraud
Insufficient peer review
0
20
40
60
80
100%

HAVE YOU EVER TRIED TO PUBLISH A REPRODUCTION ATTEMPT?

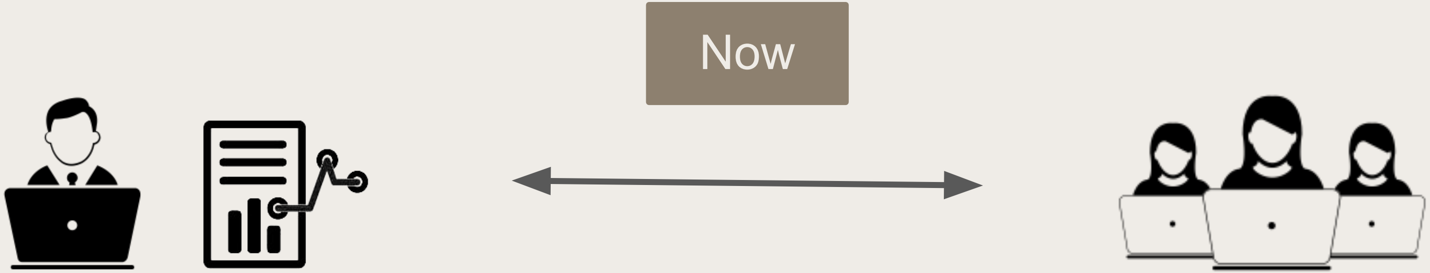
Although only a small proportion of respondents tried to publish replication attempts, many had their papers accepted.

HAVE YOU ESTABLISHED PROCEDURES FOR REPRODUCIBILITY?

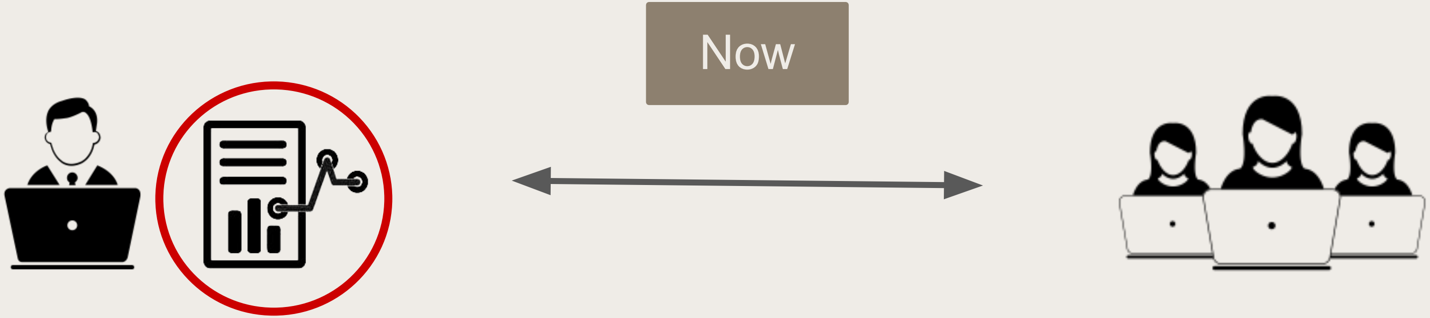
Among the most popular strategies was having different lab members redo experiments.

Number of respondents from each discipline:
Biology **703**, Chemistry **106**, Earth and environmental **95**, Medicine **203**, Physics and engineering **236**, Other **233**

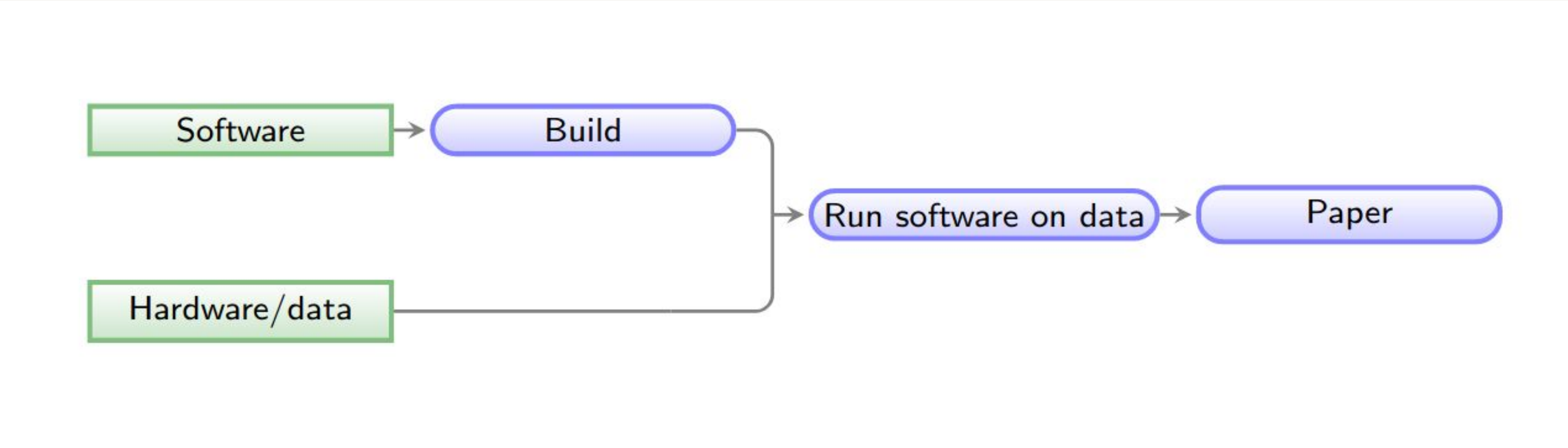
https://www.nature.com/articles/533452a



Now



Now



Software

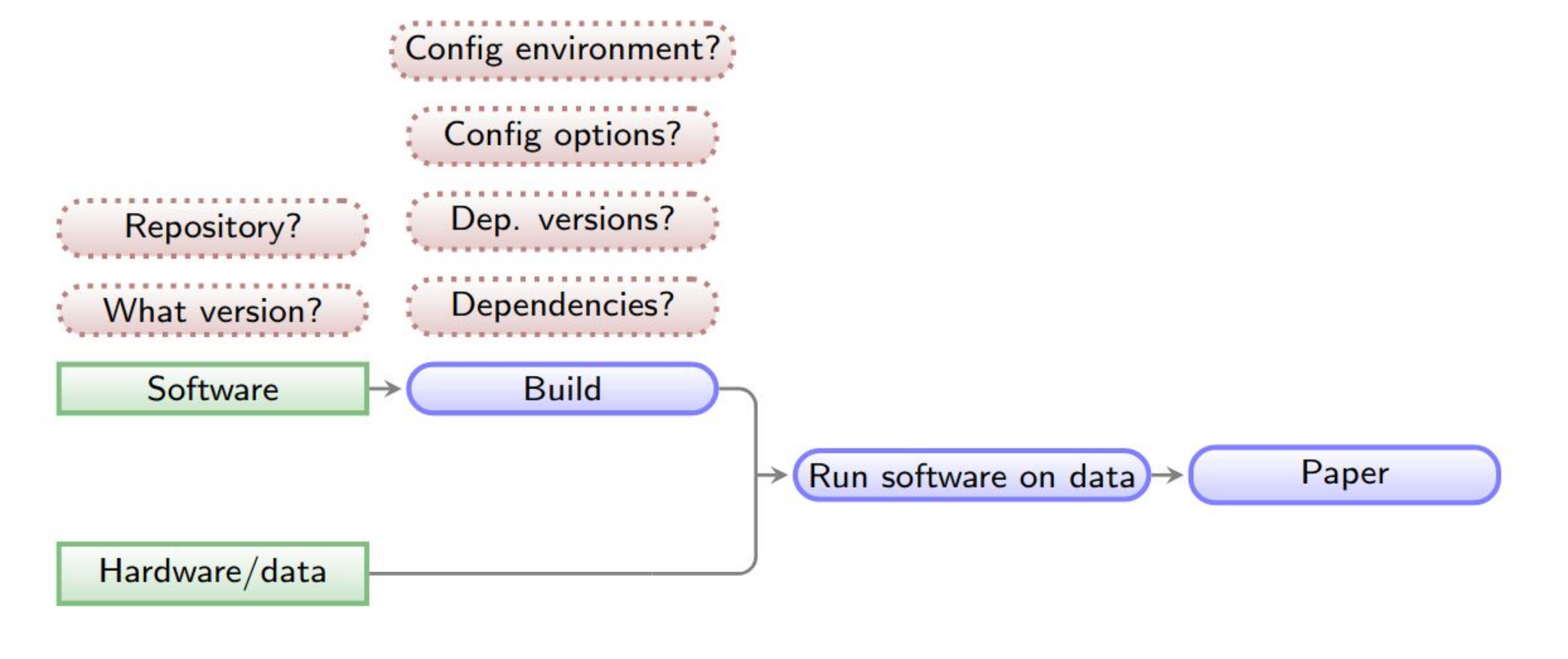
Build

Hardware/data

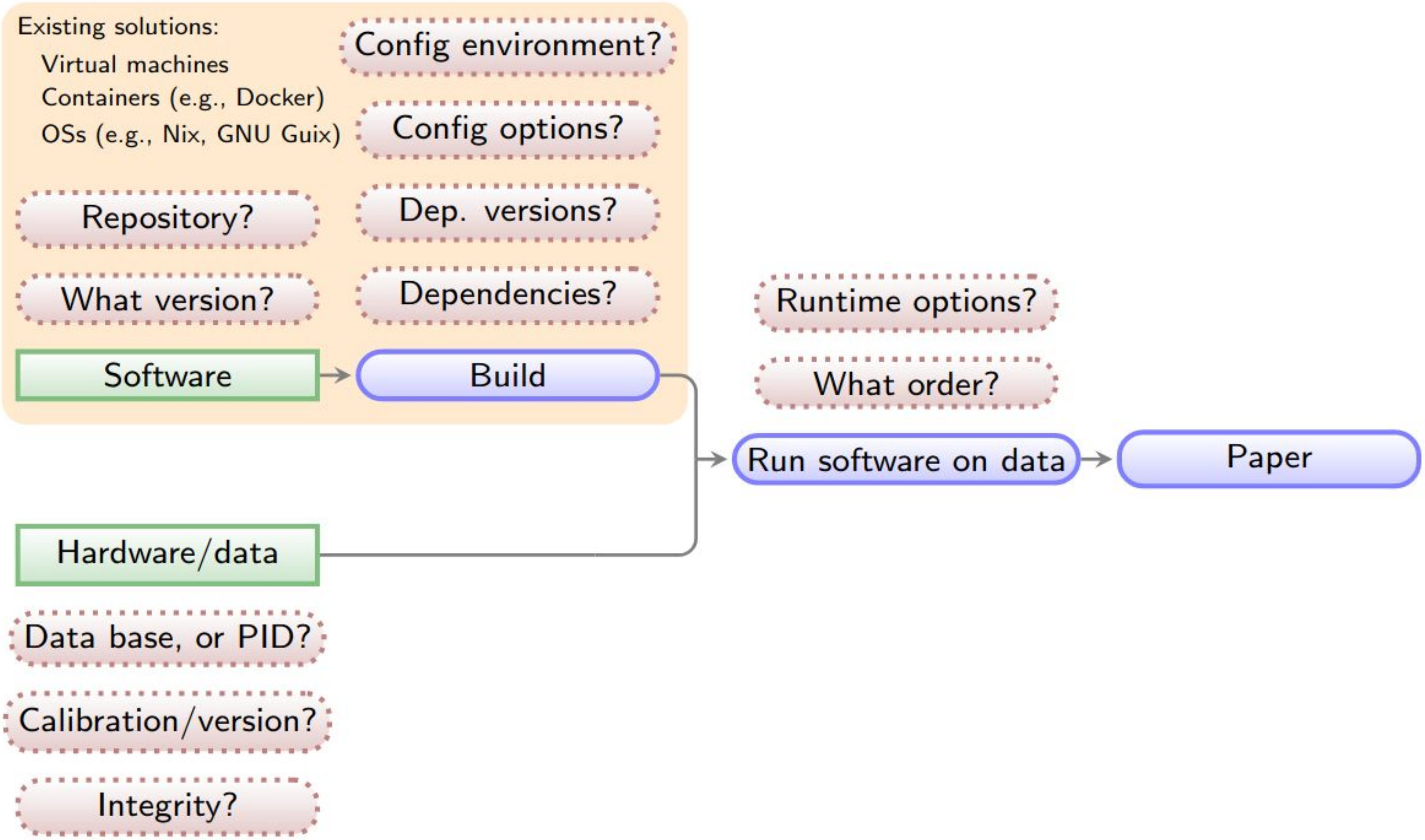
Run software on data

Paper

https://akhlaghi.org/pdf/reproducible-paper.pdf



https://akhlaghi.org/pdf/reproducible-paper.pdf



https://akhlaghi.org/pdf/reproducible-paper.pdf

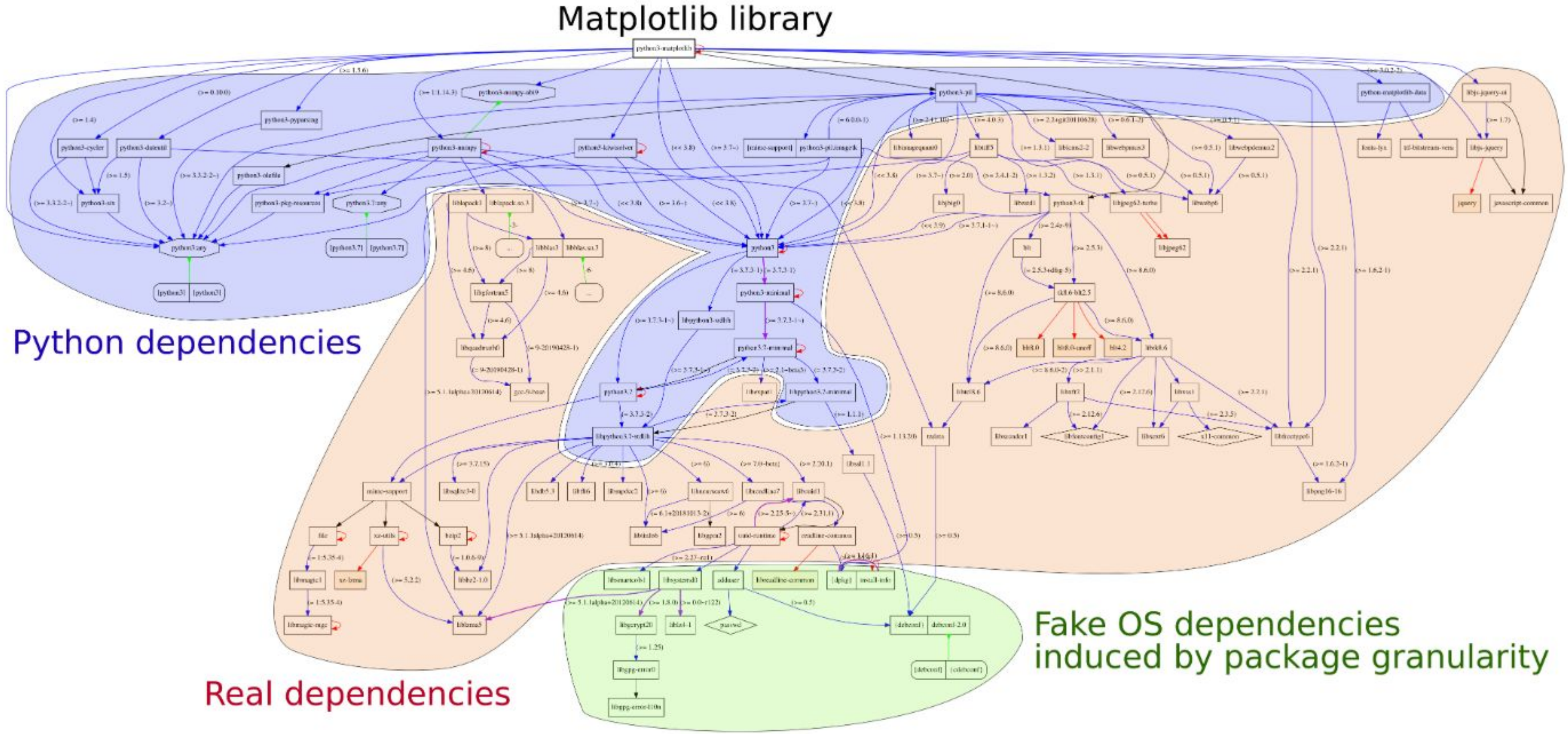


Fig. 1. Transitive dependencies of the software environment required by a simple "`import matplotlib`" command in the Python 3 interpreter.

Software dependencies

The objective is to make your analysis reproducible (by you+everyone)

An analysis may require different packages

- Load and process data
- Visualization
- Statistical analysis
- Advance processing
- …

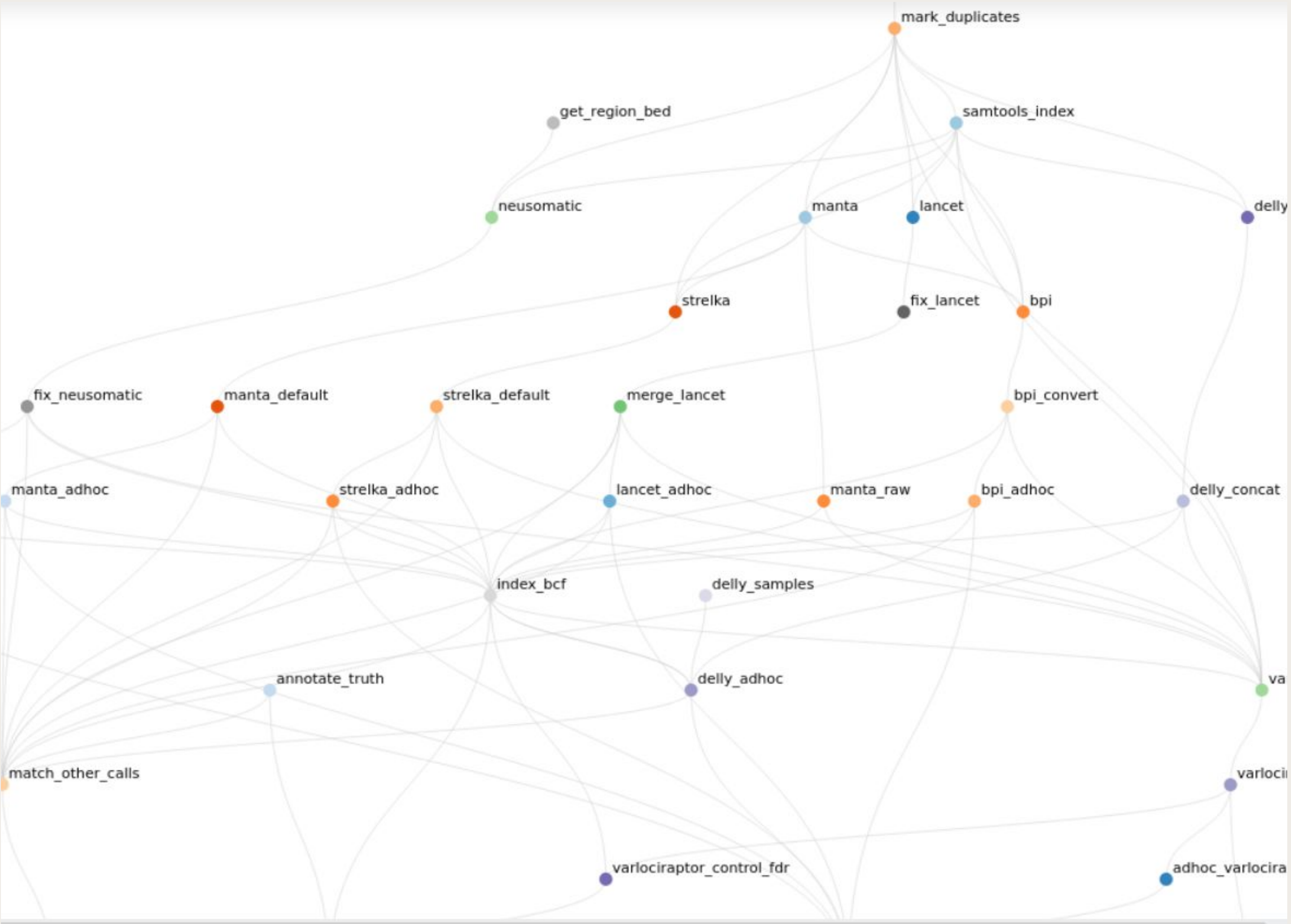
Software dependencies can change over time

An analysis that works today most probably will not work in a few years time. For example py2 → py3.

Example of reproducible workflow

Snakemake workflow

https://koesterlab.github.io/resources/report.html



How do we do it?

Tracking all the software dependencies is very hard!

Some options:

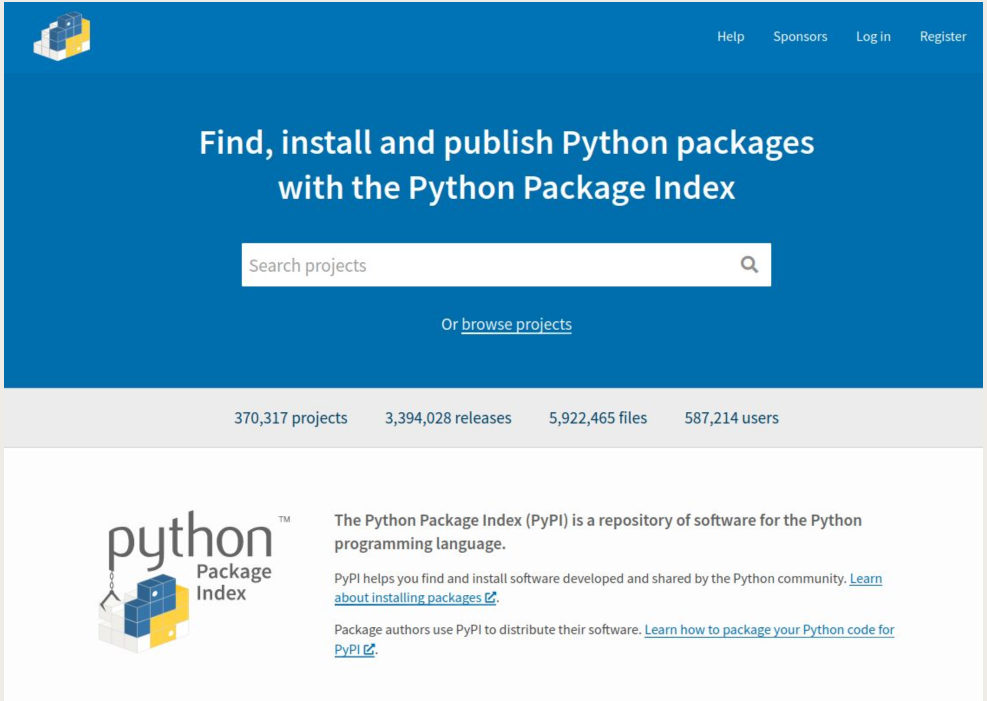
- Write in the README the list of dependencies and versions (very inefficient)
- Use a package manager (pip, Packrat, conda, …)
  - Explicitly fix the dependencies and versions
  - Incorporates a way to install them
- Containers (singularity/docker)
  - Fixed “virtual machine”, like a self-contained computer
  - Almost like a black box. Cannot be modified

pip

pip is the package installer for Python. You can use pip to install packages from the Python Package Index and other indexes.

pip

https://pypi.org/



Find, install and publish Python packages with the Python Package Index

370,317 projects 3,394,028 releases 5,922,465 files 587,214 users

The Python Package Index (PyPI) is a repository of software for the Python programming language.

PyPI helps you find and install software developed and shared by the Python community. Learn about installing packages.

Package authors use PyPI to distribute their software. Learn how to package your Python code for PyPI.

Conda / Anaconda / miniconda

Conda

What is conda?

Conda is a package manager used in scientific computing. It provides scientific libraries and dependencies.

Why conda?

- Manage the software for a project
- Can have different versions for each project
- You create virtual environments, encapsulated and reproducible

Conda

Conda

https://docs.conda.io/projects/conda/en/latest/index.html

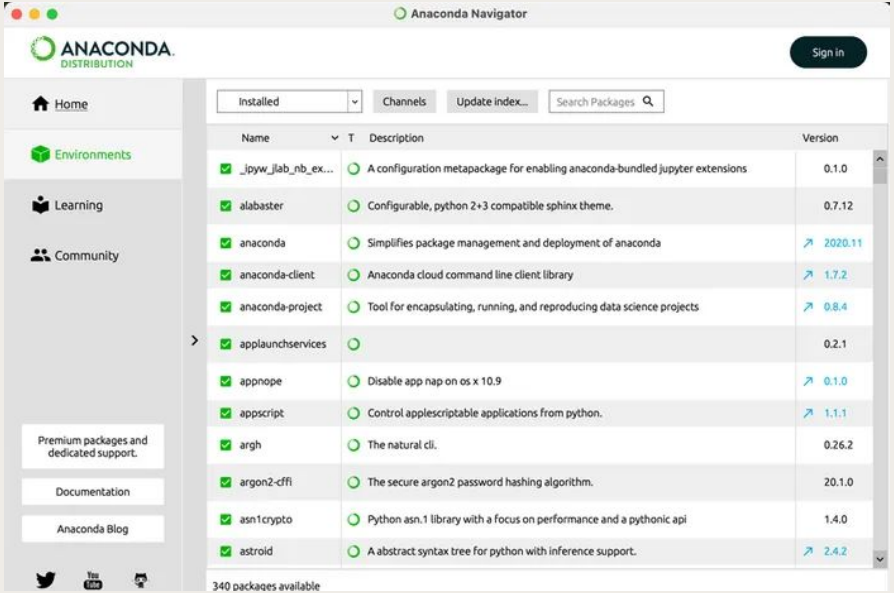
Package, dependency and environment management for any language---Python, R, Ruby, Lua, Scala, Java, JavaScript, C/ C++, FORTRAN

Anaconda

A private company providing software and package management solutions

Anaconda distribution

https://www.anaconda.com/products/distribution



	Name	T	Description	Version
☑	_ipyw_jlab_nb_ex...	○	A configuration metapackage for enabling anaconda-bundled jupyter extensions	0.1.0
☑	alabaster	○	Configurable, python 2+3 compatible sphinx theme.	0.7.12
☑	anaconda	○	Simplifies package management and deployment of anaconda	↗ 2020.11
☑	anaconda-client	○	Anaconda cloud command line client library	↗ 1.7.2
☑	anaconda-project	○	Tool for encapsulating, running, and reproducing data science projects	↗ 0.8.4
☑	applaunchservices	○		0.2.1
☑	appnope	○	Disable app nap on os x 10.9	↗ 0.1.0
☑	appscript	○	Control applescriptable applications from python.	↗ 1.1.1
☑	argh	○	The natural cli.	0.26.2
☑	argon2-cffi	○	The secure argon2 password hashing algorithm.	20.1.0
☑	asn1crypto	○	Python asn.1 library with a focus on performance and a pythonic api	1.4.0
☑	astroid	○	A abstract syntax tree for python with inference support.	↗ 2.4.2

340 packages available

conda-forge

A community-led collection of recipes, build infrastructure and distributions for the conda package manager

Community package repository

https://conda-forge.org/

https://github.com/conda-forge

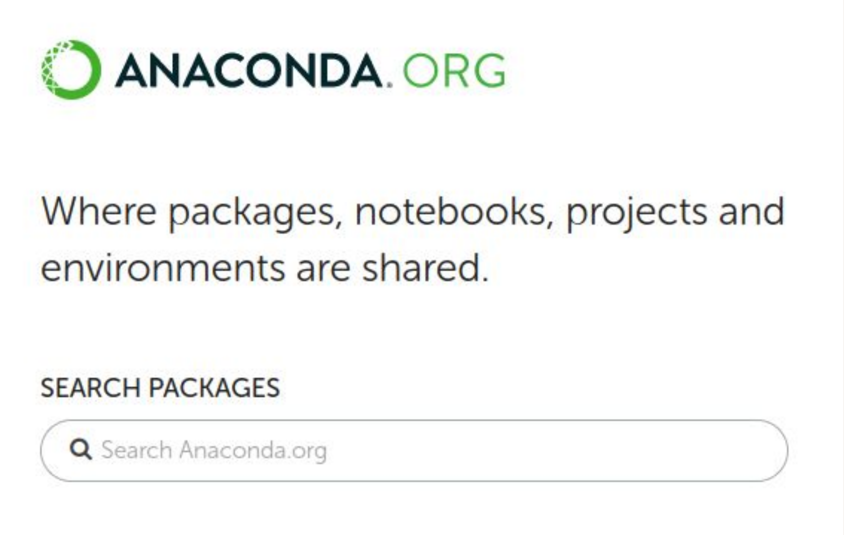
CONDA-FORGE

A community-led collection of recipes, build infrastructure and distributions for the conda package manager.

anaconda.org

Search available packages

https://anaconda.org/



ANACONDA. ORG

Where packages, notebooks, projects and environments are shared.

SEARCH PACKAGES

Search Anaconda.org

Efficiency: miniconda mamba

miniconda

A lightweight version of conda. Very easy to install

https://docs.conda.io/en/latest/miniconda.html

mamba

A very fast dependency solver. Change conda → mamba

https://anaconda.org/conda-forge/mamba

```
conda install -c conda-forge mamba
```

Demo conda

T3.1_conda